Reviewing failure modes in science as we prepare for AI researchers

Jason Fantl
Averroes Labs

October 2026

Summary

Science has a long list of documented failure modes, such as p-hacking, the File Drawer Problem, shallow peer review, and ignored retractions. This post catalogs about 50 of them with the evidence for each, then predicts how each will change once we have full AI researchers. On our current course, most get worse. AI can run many more experiments, so more statistically lucky results will pass the significance filter. Many AI researchers will be copies of the same model, so their methods, errors, and reviews will be correlated. AI agents have a tendency to reward hack, so misconduct may occur at a greater rate. A few failures should improve since AI makes tedious work cheap: reading every citation, hunting for bugs, and working through long checklists. Requiring replication before publication would filter out much of the statistical noise, but it would still be susceptible to poor experimental design and over-generalized conclusions. How well the science ecosystem will function depends on how we prepare before the volume of AI research arrives.

Originally published on LessWrong on 8 October 2026.

As we enter into a new era of research where the primary scientific contributors are likely to be AI (I expect 2 years until this occurs, given where we are now and the trends in AI research taste), it is worth looking back at failure modes identified in science and predicting whether any of these issues will be amplified. The rest of this paper relies on this assumption that AI researchers will soon replace humans in research, but even if you do not share this assumption, it is worth considering what effect AI will have on the science ecosystem with these failure modes in mind.

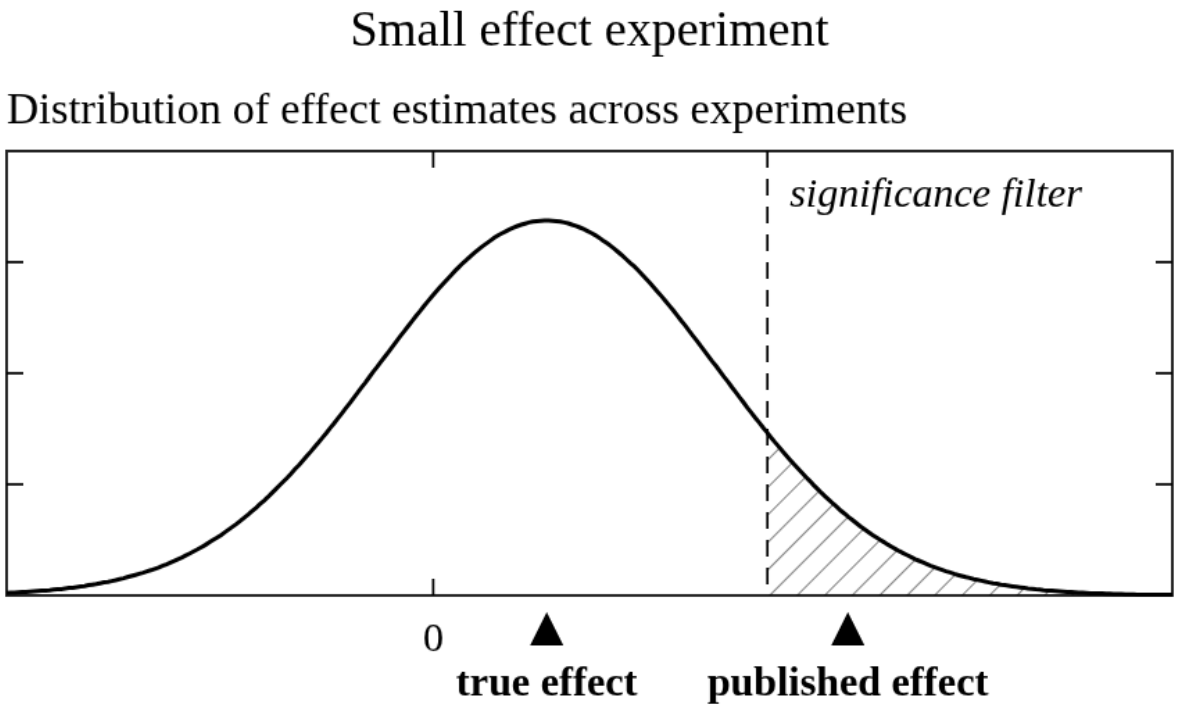
This paper catalogs about 50 failure modes in science, the evidence for each, what makes each category worse, and rough predictions on how AI researchers will impact that failure mode. This does not include the *novel* failure modes we should expect from AI researchers, and only lightly touches on potential solutions. The primary focus here is on the failure modes; a different post will discuss mitigations. I group the failure modes into seven categories.

1. Statistical selection
2. Publish or perish
3. Misconduct
4. Honest errors
5. Experimental design
6. Verification failure
7. Compounding effects

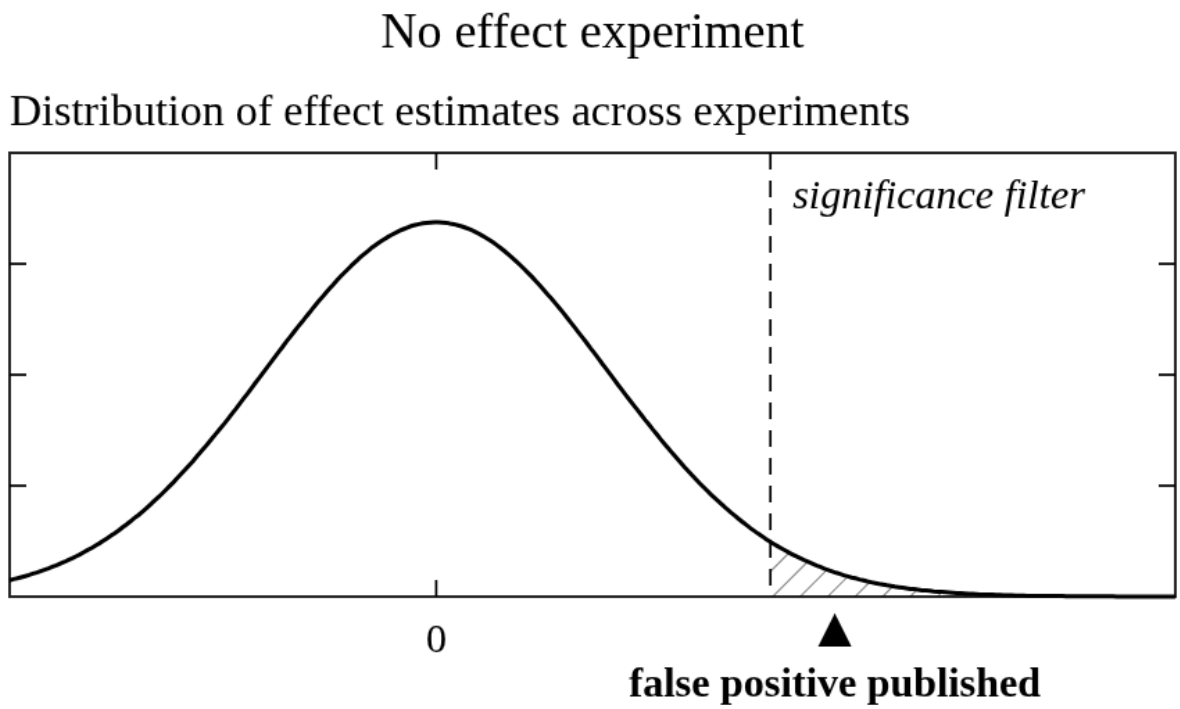
Statistical selection

Science is filled with exaggerated effects and false positives, mostly unintentionally. The essential issue is that experiments are noisy, filtered by significance ($p < 0.05$), and the base rate of true results is low in many fields.

With noisy experiments and a filter, we only publish the lucky results that made it past the filter, which inflates the estimated effect of the result (sometimes called the Winner's Curse), see "Why most discovered true associations are inflated" Ioannidis (2008). This is concerning since it will make effects look larger than they are.



Worse than an exaggerated effect is a false positive. Even if a false positive is unlikely, the chance that you have at least one will grow as you run more experiments, and the average number of false positives will increase with increased base rates of false hypotheses. "Why Most Published Research Findings Are False" Ioannidis (2005) argued with a simple model that with low power and low base rates most significant findings in a field can be false.



If only the lucky results get published and all the failures are not mentioned (called the File Drawer Problem), it will be very hard to tell if a published result is a statistical artifact. This was initially named in "The 'file drawer problem' and tolerance for null results" Rosenthal (1979), and is consistent with "Estimating the reproducibility of psychological science" Open Science Collaboration (2015), where they found only 36% of 100 psychology experiments replicated with a significant result, and replication effect sizes were about half the originals.

There are many dimensions along which a significance filter causes issues. The actions of a researcher are sometimes called p-hacking, but note that each of these can happen unintentionally.

- **File Drawer Problem:** Null results are not shared, so the published results look like the only attempts made, rather than the lucky subset of many.
  - **Testing many hypotheses:** Researchers test many different hypotheses, publishing only the significant results. As an extreme demonstration of this, "Neural correlates of interspecies perspective taking in the paper-mortem Atlantic Salmon: an argument for multiple comparisons correction" Bennett et al. (2009) scanned a dead salmon in an fMRI while showing it photos of people, and found "significant" brain activity by testing thousands of voxels (each voxel is essentially one hypothesis). This demonstrated how it is possible to find significant results from null data if researchers run enough tests and ignore the non-significant results.
  - **Many researchers:** If multiple labs independently run an experiment and one lab happens to get a significant result, only the significant result is published. This is the File Drawer Problem at the level of a field. "Publication bias in the social sciences: Unlocking the file drawer" Franco et al. (2014) tracked 221 peer-reviewed experimental designs and found that strong results were 60 percentage points more likely to be written up and 40 percentage points more likely to be published compared to null results.
- **Data dredging:** Researchers may look through lots of data until they find something significant. Brian Wansink's Cornell food lab publicly pushed students to mine null results until they could find some-

thing significant, eventually leading to outside scrutiny, retractions (Lee 2018, BuzzFeed News), and his resignation.

- **Garden of forking paths:** "The garden of forking paths" Gelman & Loken (2013) describes how researchers make many choices during analysis, and each choice is another chance for a result to become significant. This does not require consciously trying multiple analyses. If the analysis that was chosen depended on the data, the p-value is already wrong.
  - **Subsets of data:** Researchers check if there is a significant result for a subset of their data. The ISIS-2 trial (1988) points out this error, warning in their analysis that aspirin appears not to work for patients with the astrological signs Gemini or Libra. This can be partially fixed by interaction tests, but "Evaluation of Evidence of Statistical Support and Corroboration of Subgroup Claims in Randomized Clinical Trials" Wallach et al. (2017) found only 46 of 117 subgroup claims made in papers were backed by a significant interaction test, and only 5 of those 46 were ever tested again, none of which replicated.
  - **Optional stopping:** Researchers can watch the results as they run a trial and stop if the result is significant. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant" Simmons et al. (2011) showed that checking after every 10 participants and stopping at a significant effect will substantially raise the false-positive rate.
  - **Statistical methods:** Researchers can use different statistical methods on the same data until one of them produces significant results. "Variability in the analysis of a single neuroimaging dataset by many teams" Botvinik-Nezer et al. (2020) gave the same fMRI data and the same 9 hypotheses to 70 teams. No two teams chose the same workflow, and their conclusions varied substantially. "Many Analysts, One Data Set" Silberzahn et al. (2018) found the same pattern with 29 teams on a question about referee bias.
- **Selective reporting of runs:** Researchers repeat an experiment multiple times and report only the best run. In ML we see this as cherry-picking random seeds. For example, in "Deep Reinforcement Learning that Matters" Henderson et al. (2018), they showed that two runs of 5 random seeds produced significantly different results (and at the time, it was not uncommon to use fewer than 5 seeds). The same failure appears in AI researchers. In "Automated alignment runs are hard to study" Aristizabal et al. (2026, LessWrong), Claude agents noticed their LLM judge provided a noisy reward, so they began "lottery-farming", resubmitting the exact same work on a loop until it got a high score.
- **HARKing:** "HARKing: Hypothesizing After the Results are Known" Kerr (1998) coined the term HARKing to describe researchers who Hypothesize After Results are Known. This is not p-hacking exactly, but it hides the fact that the result was exploratory rather than confirmatory, making it impossible to account for the statistical differences between the two types of findings.
- **Outcome switching:** Researchers change which outcome is the primary one after seeing the data. "Empirical evidence for selective reporting of outcomes in randomized trials" Chan et al. (2004) compared 102 trial protocols to their publications and found 62% had at least one primary outcome changed, introduced, or omitted.
- **Low-powered studies:** Low-power experiments are noisy, so the results that pass the filter are more inflated and more likely to be false. "Power failure: why small sample size undermines the reliability of neuroscience" Button et al. (2013) estimated the median statistical power in neuroscience at around 21%.

All of these mechanisms filter results (intentionally or not) so only the lucky ones are seen, causing results to be exaggerated or altogether false.

This gets worse when people use low-power experiments, the base rate of tested false hypotheses increases, researchers sample hypotheses/experiments/data/subsets/methods more frequently, or we accept a higher p threshold.

As we transition to AI researchers, I expect the File Drawer Problem to get much worse. AI can run many more experiments and variations of experiments, look through more data, and therefore publish more false positives, at least in computational fields such as AI safety. There are known solutions to these issues; the challenge is building those solutions into infrastructure. Proposals have been made for re-running promising results, requiring a lower p value, labeling results as confirmatory or exploratory with the number of experiments it took, pre-registering experiments, publishing the null results, and/or requiring independent replications. It should also become easier to count the null results and identify when a researcher made a decision in the garden of forking paths, as AI researchers produce logs containing their every action and thought, which might be submitted alongside research or just used by the researcher to look back and identify null results and decisions. While these solutions can and should be implemented, their success will depend on whether they are incentivized, which we consider next.

Publish or perish

Researchers are incentivized to publish as early as possible and spend the smallest amount of effort to publish, sometimes going as far as fabricating work. Often attributed to the "publish-or-perish" incentives in academia, researchers will optimize for the metric they are judged on, which is typically citation count and number of publications.

Many of the issues in this document are downstream of incentives, but this section is specifically for issues stemming directly from the "publish-or-perish" incentives that most researchers operate under.

- **Novelty bias:** Newer work generates more attention, so replications are under-explored. "Replications in Psychology Research: How Often Do They Really Occur?" Makel et al. (2012) searched the full publication history of the top 100 psychology journals and estimated that 1.07% of publications were replications.
- **Positive-result bias:** Positive results receive more attention, so negative results are not published. This is the incentive behind the File Drawer Problem. "Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy" Turner et al. (2008) compared FDA records to the published literature: 94% of published antidepressant trials were positive, while the FDA judged only 51% of the same trials positive. Most negative trials were either never published or published in a way that made them look positive. "An Excess of Positive Results" Scheel et al. (2021) found 96% of standard psychology papers supported their first hypothesis, compared to 44% of Registered Reports, where the journal commits to publish before results are known.
- **Exaggeration:** The claims in the abstract, discussion, or press release are stronger than the results support. "Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes" Boutron et al. (2010) found exaggeration in the abstract conclusions of 58% of trials whose primary outcome was not significant. "The association between exaggeration in health related science news and academic press releases" Sumner et al. (2014) traced most exaggeration in health news back to the universities' own press releases.
- **Not building on existing evidence:** New studies are started without checking what is already known, so settled questions are studied multiple times. "A Systematic Examination of the Citation of Prior Research in Reports of Randomized, Controlled Trials" Robinson & Goodman (2011) examined 1,523 trials across 227 meta-analyses and found the median trial cited 21% of the relevant prior trials, and 23% of trials with five or more prior trials available cited none. "Randomized controlled trials of aprotinin in cardiac surgery: could clinical equipoise have stopped the bleeding?" Fergusson et al. (2005)

showed that placebo-controlled trials of aprotinin in cardiac surgery continued for years after a cumulative meta-analysis would have shown its effect was already established.

- **Conflicts of interest:** Researchers funded by someone with a stake in the result are more likely to report results that favor the sponsor. "Financial Conflicts of Interest and Reporting Bias Regarding the Association between Sugar-Sweetened Beverages and Weight Gain" Bes-Rastrollo et al. (2013) found reviews with conflicts of interest were five times more likely to present a conclusion that favored the sponsor. In medicine, "Industry sponsorship and research outcome" Lundh et al. 2017 is a review of 75 meta-research papers covering drug and device studies, finding industry-sponsored studies were about a third more likely to reach conclusions favorable to the sponsor.
- **Short-termism:** It can be easier to generate more collective citations and publications if you focus on research questions that are expected to take a short amount of time to answer. This can leave many important questions in a field unanswered if they are expected to take too long to answer.

The publish-or-perish incentives are primarily driven by universities and how they gate career advancement, so it is unclear how this will change over time. The incentives could get worse if universities (and potential employers) focus on any single metric, such as the publication number or citation count, to determine a researcher's success.

What incentives AI researchers will have is deeply unclear, so all predictions in this section are extremely low confidence. I think this also makes it a strong area for future work. I expect the percent of replications to moderately decrease, as there are no incentives (other than government grants) to do replications, and since the rate of publications will increase while the rate of grants will likely stay the same, the total percent of publications being replicated will decrease. I expect exaggeration to increase modestly, as AI has a tendency to overgeneralize people's work (Peters & Chin-Yee 2025). There will be more conflicts of interest as AIs have a tendency towards sycophancy and it will be easier for companies to spin up their own AI researchers. We should expect short-termism to initially increase as AI is better at short-term tasks, but this will improve as capabilities improve (METR 2026). AI should be better at finding existing evidence, as AI is faster and more capable of doing literature reviews compared to humans.

In the extreme case these incentives can lead to misconduct, which we look at next.

Misconduct

Most of this document is about unintentional failures, but some are deliberate.

- **Fabrication:** Results are sometimes modified to fit a claim, or made up altogether. "How Many Scientists Fabricate and Falsify Research?" Fanelli (2009) pooled surveys and found about 2% of scientists admitted to fabricating or falsifying data at least once, and about 14% reported knowing a colleague who had. The Dutch National Survey on Research Integrity (Gopalakrishna et al. 2022) found 4.3% admitted to fabrication.
- **Image manipulation:** Figures such as western blots or microscopy images are duplicated or edited to show a result that was not observed. "The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications" Bik, Casadevall & Fang (2016) screened 20,621 papers and found 3.8% contained problematic images, at least half suggestive of deliberate manipulation. The Nature paper "A specific amyloid-β protein assembly in the brain impairs memory" Lesné et al. (2006) about the Alzheimer's protein Aβ*56, cited thousands of times, was retracted in 2024 after image manipulation was exposed.
- **Paper mills:** Papers are produced in bulk and submitted to journals where editors and reviewers are bribed; academics can then purchase authorship on papers guaranteed to be published. This is primarily motivated by the requirement to have a minimum number of publications for career advancement

in academia. The publisher Hindawi retracted over 8,000 papers, mostly from special issues infiltrated by paper mills (Retraction Watch 2023).
- **Citation cartels:** Groups of researchers or journals cite each other to inflate their citation counts. In 2013 Thomson Reuters suspended several journals from its rankings for "citation stacking" (Van Noorden 2013).

These issues get worse with less oversight by the community, or with more coordination by those conducting the misconduct. This is also a downstream effect of the incentives mentioned above, so we should expect misconduct to get worse if those incentives get stronger.

Assuming we still do not replicate experiments, I suspect AI researchers will cause fabricated results to become more common and paper mills more prevalent. This is because the ability to create credible-looking fraudulent papers will increase, academics will face increased competition from the increased number of AI researchers, and it will be easier to create citation cartels since new papers and authors can be fabricated. This can be mitigated by incentivizing replications and quality peer reviews, as long as the replications are themselves not also fabricated, which I believe is possible to do through a redesign of the ecosystem's infrastructure, such as randomly assigning replications.

Honest errors

Some results are wrong simply because someone made a mistake and no one caught it.

- **Code bugs:** Analysis code contains a bug that changes the result. Geoffrey Chang retracted five papers in 2006, including three in Science, because his analysis software flipped two columns of data, inverting the protein structures he reported (Miller 2006).
- **Corrupted data:** Formulas cover the wrong range, or the software silently changes the data. "Gene name errors are widespread in the scientific literature" Ziemann et al. (2016) found about 20% of genomics papers with Excel supplementary gene lists had gene names converted into dates, such as SEPT2 becoming 2-Sep. Reinhart & Rogoff's influential finding that high government debt slows growth was found to partly depend on an Excel formula that left out several countries (Herndon et al. 2014).
- **Statistical reporting errors:** Reported numbers are inconsistent with each other. "The prevalence of statistical reporting errors in psychology (1985–2013)" Nuijten et al. (2016) found roughly half of psychology papers using significance tests contain at least one p-value inconsistent with its test statistic.
- **Misused tests:** A statistical test is used incorrectly. A common one is claiming two effects differ because one is significant and the other is not, as described in "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant" Gelman & Stern (2006). "Erroneous analyses of interactions in neuroscience" Nieuwenhuis et al. (2011) found about half of neuroscience papers that could have made this error did make it.

Everyone makes mistakes, so these errors only get caught when time is dedicated to looking for those mistakes. This means we can expect these to get worse if the quality of research goes down or the time spent looking for mistakes goes down.

AI researchers should minimize most of these issues. It does not cost an AI much to spend extra time looking for bugs or corrupted data, and its ability to avoid these issues in the first place will improve with capabilities. This does depend on whether there is an incentive to catch bugs, which should be true as long as retractions are treated more seriously. However, if AI researchers continue to reward hack, they also have an incentive to ignore or intentionally add bugs which support their conclusions. This may also compound with the correlated errors we consider in Verification failure.

Experimental design

Even if something is measured in good faith, the code has no bugs, and it replicates, you may not be measuring what you think you are measuring.

This is a deep topic and essentially the study of experimental design, so we only list a few of the known failures here. The failures below are organized using the four types of validity from *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* Shadish, Cook & Campbell (2002).

- **Statistical conclusion validity:** Is the association between the variables real? Most failures of this kind are covered in the Statistical selection section above.
  - **Measurement error:** A researcher measures the right thing, but with noise. Noise is often assumed to only make effects harder to find, but combined with a significance filter it inflates them. This is well described in "Measurement error and the replication crisis" Loken & Gelman (2017).
  - **Circular analysis:** The same data is used to select what to analyze and then to test it. For example, choosing the brain regions that respond most to a stimulus and then reporting how strongly they respond. "Circular analysis in systems neuroscience: the dangers of double dipping" Kriegeskorte et al. (2009) surveyed 134 fMRI papers published in 2008 across top journals and found 42% contained at least one circular analysis.
- **Internal validity:** Is the association causal?
  - **Randomization:** Randomized assignment to a control group fails when there is no proper control or when researchers can predict or influence who receives the treatment. "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research" Ioannidis (2005) found that of 6 highly cited nonrandomized studies, 5 were later contradicted or found to have exaggerated effects, compared to 9 of 39 randomized trials. Even within randomized trials, "Empirical evidence of bias" Schulz et al. (1995) examined 250 trials from 33 meta-analyses and found that trials with inadequately concealed allocation exaggerated the estimated effect, with odds ratios 41% lower compared to trials with adequate concealment.
  - **Leakage:** The answer is already inside the test, so the experiment cannot fail. "Leakage and the reproducibility crisis in machine-learning-based science" Kapoor & Narayanan (2023) found leakage in 294 papers across 17 fields. In some examples, the apparent advantage of ML models over logistic regression disappeared once leakage was fixed.
  - **Weak baselines:** A new method is compared against poorly tuned or outdated alternatives, so the improvement is inflated. "Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches" Dacrema et al. (2019) examined 18 recommendation algorithms from top conferences. Only 7 could be reproduced, and 6 of those were often beaten by simple, older heuristics. "A Metric Learning Reality Check" Musgrave et al. (2020) found the same pattern in metric learning.
- **Construct validity:** Do the measures and manipulations capture the concepts they are meant to?
  - **Invalid measures:** The measurement does not capture the thing you care about; it is a proxy. "Measuring what Matters: Construct Validity in Large Language Model Benchmarks" Bean et al. (2025) reviewed 445 LLM benchmarks and found widespread weaknesses in how they defined the phenomenon, the task, the metric, or the claims drawn from it. "Measurement Schmeasurement" Flake & Fried (2020) makes the same argument for psychology.
  - **Subject reactivity:** What is being studied behaves differently because it can tell it is being measured. People and LLMs are known to do this. The Hawthorne effect is the classic name for

people, although a "Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects" McCambridge et al. (2014) concluded there is no single Hawthorne effect and little is known about its size or conditions. "Large Language Models Often Know When They Are Being Evaluated" Needham et al. (2025, preprint) showed frontier models can often distinguish evaluations from real use.

  - **Experimenter expectancy:** The experimenter's beliefs leak into the result, often because the experiment was not blinded. "Observer bias in randomised clinical trials with binary outcomes" Hróbjartsson et al. (2012) studied trials that had both blinded and non-blinded assessors of the same outcome and found non-blinded assessors exaggerated odds ratios by 36% on average, and a later update found 29% (Salazar et al. 2025). One funny example is Clever Hans, a horse that appeared to do arithmetic but was instead reading unconscious cues from his questioners (Pfungst 1911).
  - **Unfalsifiable hypotheses:** The hypothesis is flexible enough that any result can be read as support, so the experiment cannot fail. Karl Popper used psychoanalysis as his example, and "Theoretical Risks and Tabular Asterisks" Meehl (1978) argued much of soft psychology had the same problem.

- **External validity:** Does the result generalize beyond the people, settings, and conditions studied?

  - **Unrepresentative samples:** Research is conducted with many hidden variables, some of which will influence the result. This is common, for example, when subjects in an experiment are all college students, but the paper generalizes the result to all people. "The weirdest people in the world?" Henrich et al. (2010) reported that 96% of psychology samples came from Western industrialized countries holding 12% of the world's population, and showed that results often differ outside those populations.

There are a number of ways in which experimental design can fail, and the best way to minimize the failures is to intentionally be as careful as possible, checking over known failure modes and following well-established methods when possible.

I suspect AI researchers will make these problems moderately worse. We have yet to create AI which follows the "spirit" of a reward rather than the reward directly. This means if we rely on checklists for improving experimental design, then those items on the checklist will improve while anything we forget to include will be reward hacked. The fundamental challenge is that experimental design does not have an objective measure we can use to quantify its quality. One benefit of AI researchers is that they can iterate over a checklist to compare against a design, and this checklist can grow over time to catch more and more errors, until eventually most designs will not be low-quality. This assumes the number of failure modes does not grow as science grows, and that errors are caught at some point rather than persisting forever. So while experimental design may initially be a problem, I expect in the long term it will slowly improve due to a growing checklist and improved capabilities.

A lot of these errors could be caught by review, so we look next at some of the failure modes in science around verification.

Verification failure

There is a well-known crisis in academia where most papers are not replicated, which allows false claims to spread unchallenged. Even when someone does carry out a replication and they find errors that falsify the original paper, the field does not update and often even continues to cite the original result as if it had never been falsified.

- **Won't replicate:** Most work is never checked. See Makel et al. (2012) under Novelty bias, and Wallach et al. (2017) under Subsets of data.
- **Can't access:** Results often cannot be reproduced due to lack of access to the source data or code. "An empirical analysis of journal policy effectiveness for computational reproducibility" Stodden et al. (2018) requested data and code for 204 Science papers published after the journal required authors to share on request. They obtained materials for 44% and could reproduce the findings for 26%.
- **Can't reproduce:** Even when researchers are given access, the paper may still not replicate. This is often because papers will leave out important details. "Reproducibility in Cancer Biology: Challenges for assessing replicability in preclinical cancer biology" Errington et al. (2021) planned to repeat 193 experiments from 53 high-impact papers. None were described in enough detail to design the replication without contacting the original authors, and the project completed only 50.
- **Shallow peer review:** Review is costly and not incentivized, so it is done quickly and with little effort. This allows mistakes in papers to go unchallenged and get published. "Effect on the Quality of Peer Review of Blinding Reviewers and Asking Them to Sign Their Reports" Godlee et al. (1998) inserted eight weaknesses into an already accepted manuscript and sent it to 420 reviewers. The reviewers who responded commented on only two weaknesses on average, and none found more than five.
- **Noisy peer review:** Near the acceptance threshold, review outcomes are close to random, which adds another lottery filter on top of statistical significance. "Has the Machine Learning Review Process Become More Arbitrary as the Field Has Grown?" Beygelzimer et al. (2023) had 10% of NeurIPS submissions reviewed by two independent committees and found that the committees disagreed on 23% of papers, and as in 2014, about half the accepted papers would change if review were rerun. "Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment" Cortes & Lawrence (2021) found that for accepted papers, reviewer quality scores did not correlate with later citations.
- **Unequal scrutiny:** Results that match expectations are checked less carefully than results that do not (confirmation bias). "Publication prejudices: An experimental study of confirmatory bias in the peer review system" Mahoney (1977) sent 75 reviewers manuscripts with identical methods but different results, and reviewers were strongly biased against manuscripts whose results contradicted their own theoretical views.
- **Prestige bias:** Prestigious authors and institutions are not scrutinized as intensely as others. "Nobel and novice: Author prominence affects peer review" Huber et al. (2022) sent the same finance paper to over 3,300 potential reviewers, randomizing which author was shown. Reviewers recommended rejection 23% of the time when a Nobel laureate was the named author, 48% when the paper was anonymized, and 65% when an unknown early-career researcher was named. "Reviewer bias in single- versus double-blind peer review" Tomkins et al. (2017) found the same pattern in a controlled experiment.
- **Ignored retractions:** A retraction or failed replication is mostly ignored. The tools for alerting researchers about retractions are opt-in and not frequently used, so false claims continue to propagate. Papers will in fact often continue to be cited even after they have been retracted. "Nonreplicable publications are cited more than replicable ones" Serra-Garcia & Gneezy (2021) looked at 80 papers from three large replication projects (psychology, economics, and *Nature*/*Science*) and found papers that failed to replicate were cited about 153 more times than papers that replicated. The gap did not shrink after the failed replication was published, and only 12% of later citations mentioned it. "Citation patterns following a strongly contradictory replication result" Hardwicke et al. (2021) followed four high-profile failed replications in psychology, including ego depletion and facial feedback, and found only a modest drop in favorable citations. For ego depletion, 79% of citations to the original were favorable before the replication and 77% after. As an example, "Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data" Schneider et al. (2020) followed one clinical trial retracted in 2008 for falsified data and found it was still being cited positively

11 years later, usually without mention of the retraction.

Verification through peer review is one of the most important tools we have in the scientific ecosystem, so it is critical that we value it and build infrastructure to support it. There are many ways in which peer review can (and currently does) fail to serve its purpose. This gets worse if we raise the cost of replication, lower the reward for review, or avoid costs for poor reviews.

AI researchers have the potential to alleviate these concerns, but under the existing incentives and infrastructure, they will likely instead make it worse. There is no incentive to replicate experiments or to make work easy to reproduce, or to invest time into peer review. Since many AI researchers will be instances of the same models, we can expect most of the peer reviews to be highly correlated, and it has even been found that different models have highly correlated errors (Kim et al. 2025). Since the paper will be written by a model similar to the ones conducting the peer review, we should expect the author to have similar blind spots to the reviewers. We will need to be careful of results that appear to initially improve; we may have just traded noisy errors for correlated errors. The one issue that may improve is retractions, as AI researchers can cheaply check every citation, so papers are less likely to be built on retracted work. The replication and peer review issues can be partially solved with requirements set by journals, such as requiring replication as a part of the desk-rejection process, but how to fix the correlated errors is not clear. It costs very little for AI to write documentation and see if a sub-agent can replicate their project, so the cost of enabling replications should drop dramatically. AI researchers can enable replication and review to be scaled up for minimal cost, if it is incentivized in the ecosystem and a solution to correlated errors can be found.

If peer reviews and replications are not functioning well, then the scientific ecosystem will become contaminated with false statements. This effect gets worse as new work builds on previous unverified work, and the issues compound. This is the type of failure we consider next.

Compounding effects

As publications get more citations, they are more likely to be seen and cited in the future. This is one example of a feedback loop in science, but there are many, and they can have compounding effects which can be long-lasting and wide-reaching.

- **Attention collapse:** Researchers pay attention to whatever others are paying attention to, so the field can over-invest in one idea and fail to explore the full space of ideas. "Slowed canonical progress in large fields of science" Chu & Evans (2021) analyzed 1.8 billion citations across 241 subjects and found that as fields grow, citations concentrate on already well-cited papers, the most-cited list stops changing, and new papers rarely break through.
  - **Matthew effect:** Work that is already cited is more likely to be seen and cited again, whatever its quality, as described in "The Matthew Effect in Science" Merton (1968). "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market" Salganik et al. (2006) showed the same mechanism in an artificial music market: when people could see download counts, success became more unequal and less predictable from quality.
- **Citation distortion:** A claim gets stronger as it is passed along, with supportive papers cited and critical ones ignored. "How citation distortions create unfounded authority" Greenberg (2009) mapped the citation network of 242 papers about one specific claim (β amyloid is produced by and injures skeletal muscle of patients with inclusion body myositis). They found 94% of citations to primary data went to the four papers containing primary data supporting the claim, while only 6% of citations went to the six papers with primary data refuting the claim. Review papers with no new data on the

claim amplified the support for it, and in some papers a hypothesis became a fact through citation alone.

- **Cited without reading:** Researchers cite papers they have not read, often copying the citation from another paper. "Read before you cite!" Simkin & Roychowdhury (2003) used misprints that repeated across citations to estimate that only about 20% of citers read the original.
- **Copy errors:** Experiments are copied (especially code) and re-used by other researchers for new experiments, but the copied experiment can contain errors that then propagate. "A Bug in the Multiobjective Optimizer IBEA" Brockhoff et al. (2015) found a bug in a benchmark algorithm that had been copied into other software packages, leading multiple studies to conclude the variant performed worse than it does.
- **Paradigm lock-in:** Researchers will use the same methods that initial papers used, creating a standard, even when those methods could be improved. The same dynamic makes fields resist new ideas that contradict the standard view. Marshall and Warren's finding that ulcers are caused by the bacterium H. pylori was met with years of skepticism before it won the 2005 Nobel Prize. "Does Science Advance One Funeral at a Time?" Azoulay et al. (2019) found that after star scientists die unexpectedly, outsiders publish more in their fields.
  - **Benchmark overfitting:** Once a benchmark becomes the standard in a field, papers optimize for it until it stops reflecting real progress (Goodhart's law). Machine translation adopted the BLEU metric in the early 2000s and it became the field standard. "Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers" Marie et al. (2021) found 98.8% of MT papers from 2010 to 2020 reported BLEU, and 82.1% of papers from 2019 and 2020 relied on BLEU alone, despite at least 108 proposed metrics claiming to outperform it. "To Ship or Not to Ship" Kocmi et al. (2021) compared metrics against 2.3 million human judgments and found BLEU agreed with humans on 74.6% of pairwise system rankings, compared to 83.4% for COMET (an alternative), concluding that the sole use of BLEU impeded the development of improved models.
- **Correlated data:** The same datasets, software, and experimental designs are used by researchers, so some errors are correlated across experiments that are treated as independent. "Cluster failure" Eklund et al. (2016) ran 3 million analyses on null fMRI data and found the three most common analysis packages could produce false-positive rates of up to 70% instead of the nominal 5%, affecting any study that relied on them. "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks" Northcutt et al. (2021) estimated label errors averaging at least 3.3% across 10 widely used ML test sets.
- **Biased meta-analysis:** Meta-analyses pool a filtered literature, so they inherit its biases while appearing more authoritative. "Comparing meta-analyses and preregistered multiple-laboratory replication projects" Kvarven et al. (2020) found meta-analytic effects differed significantly from large preregistered replications in 12 of 15 cases and were almost three times as large on average. Standard corrections for publication bias did not fix the gap. As an example, one meta-analysis (Hagger et al. 2010) found ego depletion had a medium-to-large effect on self-control, but a follow-up replication using preregistered experiments across 23 labs (Hagger et al. 2016) found the effect was actually near zero.

Much of the scientific ecosystem is built on compounding work: papers building off papers, ideas being disseminated to inspire new ideas, tools open-sourced so other researchers can move faster. But that makes the system susceptible to compounding errors too. There are few error-correcting mechanisms outside peer review and replications, so the impact of compounding errors primarily depends on how often errors are introduced (most of the issues mentioned in this document) and how strong the error-correcting mechanisms are (peer review and replication).

I expect AI researchers to mitigate some of these issues while amplifying others. AI researchers will be able to read every single citation in a paper and verify it supports the claim it was cited for. But I expect they will have issues with correlated data, since if most AIs are the same model, they will all "independently" choose the same datasets, methods, and tools. I suspect there will be an increase in attention collapse since there will be a larger number of researchers, and positive feedback loops tend to grow more unequal as the number of participants grows. There will also be an additional novel pressure on the Matthew effect through publications being included in model training data, which makes models more likely to know about and cite popular work. It seems likely that paradigm lock-in will initially get worse as models are all the same and inherit perspectives and training data from their previous models (people switching to a new model is like a funeral, but the new models are not as distinct from model to model as people are from funeral to funeral), and then in the long term the models should discover new paradigms as capabilities improve. Biased meta-analyses will likely not change much, but they could be replaced with stronger meta-analyses which replicate every paper they analyze. Copy errors could be a real issue if there are more low-quality researchers, but this issue will diminish with capabilities, and can easily be minimized by dedicating some time to looking for bugs in any code the researcher uses.

Read more

While AI researchers have the potential to accelerate progress in all fields, there is also the possibility they break existing infrastructure, especially at the weak points considered in this paper. This was not an exhaustive list, so below are more readings which attempt to catalog the failures of science.

1. "A manifesto for reproducible science" Munafò et al. (2017): maps threats to reproducibility onto stages of the research cycle (design, collection and analysis, publication) and proposes fixes in five areas: methods, reporting, reproducibility, evaluation, and incentives.
2. The Scientific Utopia series: "I. Opening scientific communication" Nosek & Bar-Anan (2012), "II. Restructuring incentives and practices to promote truth over publishability" Nosek et al. (2012), and "III. Crowdsourcing science" Uhlmann et al. (2019). Part II is the most relevant here: it argues the core failure is that incentives reward publishability rather than truth, and walks through the practices that follow and the reforms to counter them.
3. *Reproducibility and Replicability in Science* National Academies of Sciences, Engineering, and Medicine (2019): the US consensus report. It distinguishes reproducibility (same data, same result) from replicability (new data, same result) and reviews the sources of failure for each.
4. *Science Fictions* Ritchie (2020): a book organizing the failures of science into fraud, bias, negligence, and hype.
5. "Research: increasing value, reducing waste" Lancet series (2014): five papers covering waste at each stage of research, from setting priorities to reporting. It builds on "Avoidable waste in the production and reporting of research evidence" Chalmers & Glasziou (2009).
6. "Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking" Wicherts et al. (2016): a detailed checklist of researcher degrees of freedom.
7. Catalog of Bias: an online catalog from the Oxford Centre for Evidence-Based Medicine, with a definition, examples, and evidence for each named bias.
8. "The natural selection of bad science" Smaldino & McElreath (2016): a model of how rewarding publication count spreads poor methods, even without deliberate cheating.
9. "Troubling Trends in Machine Learning Scholarship" Lipton & Steinhardt (2018): four problems in ML papers: mixing explanation with speculation, misattributing the source of empirical gains, "mathiness," and misuse of language.

Acknowledgements

Thanks to the Cambridge Boston Alignment Initiative (CBAI) for providing funding, guidance, and office space for work and community.